GIE®

GENERAL INFORMATION EVALUATION

Technical and Interpretive Manual

General Information Index (GII)

DOCUMENT VERSION	2026.1
LAST UPDATED	July 2026
TEST VERSION	GIE v2024.1
TEST AUTHOR	Xavier Jouve (2024)
MANUAL AUTHOR	Xavier Jouve, Head of Research
CONTRIBUTING EDITOR	Priya Sharma, Head of Assessment Development

CONTENTS

1 Introduction and Theoretical Foundation
2 Description of the Test and Its Scores
3 Development Status
4 Administration
5 Scoring and Equating

6 Reliability and Precision
7 Validity Evidence
8 Score Interpretation
9 Appropriate Use and Limitations
— References

CHAPTER 1

Introduction and Theoretical Foundation

Acquired general knowledge, the breadth of factual information a person has absorbed and retained, is among the most stable of the broad cognitive abilities and one of the most predictive of real-world attainment. In the Cattell-Horn-Carroll (CHC) taxonomy it sits within crystallized intelligence (Gc) as the narrow ability *general (verbal) information*, Ko (McGrew, 2009). The construct traces to the original fluid–crystallized distinction (Cattell, 1963), was consolidated by the three-stratum factor-analytic synthesis (Carroll, 1993), and is carried forward by the contemporary CHC framework that organizes most current intellectual assessment (Schneider & McGrew, 2018).

The reference instrument for this ability in individual assessment is the Information subtest of the Wechsler Adult Intelligence Scale (WAIS-IN; Wechsler, 2008), a short set of examiner-administered general-knowledge questions reported on the Wechsler subtest metric. The General Information Evaluation (GIE; Jouve, 2024) measures the same construct at much greater length: a fixed form of 157 dichotomously scored items spanning history, science, geography, the arts, and current events, self-administered online without a time limit, and reported as the General Information Index (GII) on the conventional deviation metric (M = 100, SD = 15). Three design commitments define the instrument: content samples the knowledge domains that carry the construct rather than any single curriculum; administration is untimed, so that speed contributes no construct-irrelevant variance; and the

reporting scale is linked to the WAIS-IN criterion, so that results can be read directly on a metric practitioners already interpret.

The claim that the GIE and the reference subtest read one dimension is documented rather than assumed. In a dedicated linking study (Jouve, 2026b), a normal-ogive item factor analysis of the combined item pool recovered a single latent dimension almost exactly (residual RMS = .011; Tanaka goodness-of-fit index = .999), and on a common Rasch scale the WAIS-IN items fall entirely within the difficulty range the GIE covers, the GIE extending the continuum above the subtest's hardest item. Total scores on the two instruments correlate .82 in the linked sample, and the 157-item form measures with high internal consistency (KR-20 = .970; $\xi_{\text{JRRE}} = .970$). Chapters 6 and 7 present this evidence in full.

This manual is organized in the sequence recommended for technical documentation by the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014). Chapters 2 through 4 describe the test, its development status, and its administration; chapter 5 documents the scoring model and the equating that carries scores onto the WAIS-IN metric; chapters 6 and 7 present the reliability and validity evidence; and chapters 8 and 9 address interpretation and appropriate use. Throughout, obtained scores are treated as estimates: the standard error of measurement accompanies every precision claim, and interpretive guidance is stated in terms of confidence bands rather than points.

CHAPTER 2

Description of the Test and Its Scores

The GIE presents 157 general-knowledge items in a fixed order through the online delivery platform. Items are scored dichotomously and calibrated under a two-parameter logistic (2PL) item response theory (IRT) model; a typical administration takes 45–60 minutes. Every administration yields one primary score, the GII, and several derived expressions of the same standing: a WAIS-IN Scaled Score (1–19), a T-score, a percentile rank, a stanine, and the underlying ability estimate $\hat{\theta}$.

Table 2.1 summarizes the test's specifications for the current fixed form and for the computerized adaptive test (CAT) under development.

Table 2.1

Test specifications: fixed form and adaptive roadmap

Specification	Fixed form	Adaptive form
Construct	Acquired general knowledge (CHC *Gc*, narrow ability K0)	Same
Item format	157 dichotomous items; 2PL calibration	Adaptive pool; 2PL calibration
Content domains	History, science, geography, arts, current events	Same; targeted across ability
Administration	Online, untimed; 45–60 minutes	Online, untimed; shorter on average
Delivery	Web platform; desktop or laptop	Same
Primary score	GII (*M* = 100, *SD* = 15)	Same
Derived scores	Scaled Score, *T*, percentile, stanine	Same
Score reported from	Number correct	Response pattern

Note. GII = General Information Index; CHC = Cattell–Horn–Carroll; Gc = crystallized intelligence; K0 = general (verbal) information; 2PL = two-parameter logistic item response model; WAIS-IN = Information subtest of the Wechsler Adult Intelligence Scale (Wechsler, 2008). Derived scores are defined in Table 5.2. The adaptive column describes planned characteristics of a form still in development (Table 3.1); the difference in the last row is the whole of the difference between the two, and is explained in chapter 5.

The GII is the score on which interpretation rests; the remaining metrics restate it in alternative units for particular audiences. Because the scale is equated to the WAIS Information reference (chapter 5), one Scaled Score point corresponds to 5 GII points, and a GIE result can be read on the Wechsler subtest metric without further conversion.

Two entries in Table 2.1 carry more weight than their brevity suggests. The first is the item count. A 157-item form is long for an ability measure, and length is the source of both the precision reported in chapter 6 and the 45- to 60-minute administration that limits where the instrument can be used. The second is the last row. The fixed form reports from the number of items answered correctly, the adaptive form from the pattern of responses, and that single difference in what the score is computed from is the whole of the difference between the two forms; chapter 5 explains why the fixed form may ignore the pattern, and why the adaptive form may not. Everything else in the table is held constant across them by design, so that a score means the same thing whichever form produced it.

CHAPTER 3

Development Status

Development has proceeded in stages: a pilot phase in which 2,048 participants were tested; assembly of the 157-item fixed form; 2PL calibration of the full item set; linking of the calibrated scale to the WAIS Information criterion; and reliability analysis on the linked sample. Convergent validity work on that sample is complete, while additional criterion and subgroup studies, together with the adaptive delivery system, remain in progress. Table 3.1 records the status of each phase.

Table 3.1

Development timeline

Phase	Status	Sample / Notes
Pilot study	Complete	*N* = 2,048
Fixed-form assembly	Complete	157 items
Item calibration (2PL)	Complete	157 items
Linking to WAIS-IN	Complete	*N* = 482; moment-anchored equating

Phase	Status	Sample / Notes
Reliability analysis	Complete	*N* = 482; KR-20 and JRRE
Validity studies	Ongoing	Convergent complete (*N* = 482)
Adaptive form	In progress	—

Note. N = sample size for the phase. 2PL = two-parameter logistic; KR-20 = Kuder–Richardson formula 20; JRRE = Jouve's Randomized Reliability Estimation; WAIS-IN = Information subtest of the Wechsler Adult Intelligence Scale. The linking, reliability, and convergent-validity phases used the same linked sample of N = 482 complete cases; the published dimensionality and equating study analyzed 483 adults who completed both instruments (Jouve, 2026b). Studies of additional criteria and of subgroup invariance are in progress; none is reported in this edition.

The linking phase is documented in a citable study (Jouve, 2026b) whose structural and criterion findings anchor chapter 7. In that sample, GIE raw scores averaged 80 of 157 (*SD* = 29, range 5–152), so the fixed form was exercised across nearly its entire score range during calibration. A calibration that never observes the extremes of a scale cannot speak to them; this one very nearly did, which is what licenses the conversion of chapter 5 across the whole reported range.

Table 3.1 also records what is not yet done, and the distinction matters to a reader deciding whether the instrument fits a purpose. Reliability, dimensionality, and convergent validity are complete and are reported here in full. Studies of additional external criteria, of subgroup invariance, and of stability over time are not, and no evidence for them appears in this manual because none exists to report; chapter 9 states the interpretive consequences. The adaptive form is under development and is described in this edition only as a design commitment, never as a validated alternative to the fixed form.

CHAPTER 4

Administration

The GIE is self-administered online in a single session. The 157 items are presented in a fixed order, no time limit applies, and scoring is

automatic upon completion; most examinees finish in 45–60 minutes. Consistent with professional testing standards (AERA, APA, & NCME, 2014), the examinee should work in a quiet environment free of interruption, on a desktop or laptop computer with a stable internet connection.

Although administration is untimed, examinees benefit from working at a steady pace rather than deliberating at length on individual items; incorrect answers carry no penalty beyond the item itself. An item left unanswered scores zero and is counted among the 157 (Table 5.2), which is the conservative treatment: an examinee who abandons the form partway through is scored as though the remaining items had been failed. Completing the form therefore matters, and a score obtained from a partial administration is not a GIE score.

Because the test samples learned knowledge and is delivered without proctoring, the conditions of administration bear directly on interpretation. General-knowledge questions are, unlike a matrix or a digit span, answerable by consulting a source; an unproctored administration cannot exclude that possibility, and no statistical procedure recovers from it. For consequential decisions a quiet, supervised setting is therefore preferred and any irregularity should be recorded. Chapter 9 returns to the mode difference between the GIE, taken online and unsupervised, and the criterion against which it was linked, taken under examiner proctoring.

CHAPTER 5

Scoring and Equating

Scoring rests on three facts, each established elsewhere in this manual and each documented in the linking study (Jouve, 2026b). The items are calibrated, so that responses can be referred to a latent ability continuum. The calibration that carries the GIE and its criterion onto one scale is a Rasch calibration, under which the number of items answered correctly is a sufficient statistic for that ability. And the ability continuum is anchored to the WAIS Information metric by a linear equating. Together these make the operational conversion a single

straight line on the raw score, which is what the scoring service computes. This chapter derives that line, records the reporting rules applied to its output, and sets out the alternative conversions that were examined and set aside.

Item Scoring and Calibration

Responses are scored dichotomously (0/1). The operational bank is calibrated under the two-parameter logistic model (Birnbaum, 1968; Embretson & Reise, 2000), in which each item carries a discrimination parameter a_i and a difficulty parameter b_i:

$$P(X_i = 1 \mid \theta) \;=\; \operatorname{logit}^{-1}\!\left[a_i(\theta - b_i)\right] \;=\; \frac{1}{1 + \exp\{-a_i(\theta - b_i)\}}\,.$$

The discrimination estimates of Table 7.3 describe how sharply each item separates neighbouring abilities, and they govern item selection under the adaptive form. They do not enter the fixed form's score. For the linking, the GIE and WAIS-IN items were calibrated *concurrently* under the Rasch model by marginal maximum likelihood, which places both sets on one logit scale (Rasch, 1960; Jouve, 2026b). Chapter 7 reports the evidence that one scale is enough to hold them.

The Raw Score as a Sufficient Statistic

The Rasch model has a property no other item response model shares: the number of items answered correctly carries all the information the response pattern holds about ability (Rasch, 1960; Embretson & Reise, 2000). Two examinees who answer the same number of items correctly receive the same ability estimate, whichever items they happened to answer. Sufficiency is what licenses a conversion that reads the raw score alone: the estimate $\hat{\theta}$ is a strictly increasing function of the number correct, so a straight line on $\hat{\theta}$ is a straight line on the raw score. The scoring service exploits this directly, and no likelihood is evaluated at administration time.

Equating to the WAIS-IN Metric

So that GIE results can be read on a familiar reference scale, ability is carried onto the WAIS Information criterion (Wechsler, 2008) by a moment-anchored linear equating: the linking cohort's mean and standard deviation on the criterion are applied to the GIE's item-level ability, and both ends of the resulting scale are placed from the GIE's own item statistics (Jouve, 2026b). The WAIS-IN is natively reported as Scaled Scores (M = 10, SD = 3); those are re-expressed on the 100–15 deviation scale by $\mathrm{IQ} = 100 + 15(\mathrm{SS} - 10)/3$, so that the equating line reads

$$y(\theta) \;=\; \alpha \;+\; \beta\,\theta, \qquad \alpha \approx 113.331, \qquad \beta \approx 15.049.$$

The two coefficients are not free parameters of a curve fit. The intercept is the cohort's mean performance on the criterion and the slope its standard deviation, each re-expressed on the deviation metric; the line is fixed the moment those two moments are known. The intercept sits above 100 because the linking cohort outscored the WAIS normative sample, averaging a Scaled Score of 12.7 (SD = 3.0) where the norm is 10. That is a property of the sample, discussed in chapter 9, and not of the GIE. The function returns values directly on the 100–15 deviation scale on which the GII is reported: the GII and the deviation-IQ scale are one metric, not two. It was fitted on N = 482 linked cases spanning $\theta \in [-3.0649,\ 3.3534]$. Because the 157-item GIE and the 25-item WAIS Information subtest differ substantially in length and reliability, the mapping is properly a score *linking* rather than a strict equating of interchangeable forms (Kolen & Brennan, 2014). Table 5.1 records the fitted configuration.

Table 5.1

Parameters of the moment-anchored equating function

Element	Value
Intercept α	113.3306724

Element	Value
Slope β	15.04878838
Ability coverage, θ	−3.0649 to 3.3534
Cases requiring extrapolation	0%
Linked sample	*N* = 482

Note. α = intercept and β = slope of the equating line $y(\theta) = \alpha + \beta\theta$ on the GII metric, reported at the precision at which they are held in the scoring service; they are the linking cohort's criterion mean and standard deviation re-expressed on that metric. Ability coverage is the range of $\hat{\theta}$ observed in the linked sample (N = 482); no case in that sample fell outside it (Jouve, 2026b).

The Operational Conversion

Composing the two results, sufficiency of the raw score and linearity of the equating, collapses the whole pipeline into one line. Writing X for the number of the 157 items answered correctly, the scoring service computes

$$\widehat{\mathrm{GII}} \;=\; 0.5085\,X \;+\; 69.836, \qquad X \in \{0, 1, \ldots, 157\}, \qquad X = 157 \;\Rightarrow\; 149.7 \;\rightarrow\; 150.$$

The slope is the equating slope divided by the raw-score standard deviation of the linking cohort, and the intercept places that line through the cohort's two anchoring moments; the coefficients are therefore the same two facts as α and β, expressed in raw-score units rather than in ability units. Nothing is estimated at administration time. The function is strictly increasing, so score order is preserved across the whole range, and it is not truncated at the top: the maximum raw score converts to a GII of about 150, above the 145 that a Scaled Score of 19 denotes. That the GIE can report a standing the criterion subtest's scale cannot express is the practical consequence of its longer and harder item set, and chapter 8 returns to it.

Reported Score Metrics

The GII is the score on which interpretation rests. The remaining reported values restate it in the units a particular audience reads, and each is a fixed transformation of it. Nothing in Table 5.2 adds information to the GII, and a disagreement between two of these metrics for one examinee would be a rounding artifact, not a finding.

Table 5.2

Reported score metrics and their derivation

Metric	Derivation	Reporting rule
General Information Index	$0.5085\,X + 69.836$	Nearest integer; not truncated
WAIS-IN Scaled Score	$10 + 3z$	Nearest integer, clipped to [1, 19]
***T* score**	$50 + 10z$	Nearest integer
Percentile rank	$100\,\Phi(z)$	One decimal; bounded at <0.1 and >99.9
Stanine	$\mathrm{clip}\big(\mathrm{round}(5 + 2z),\,1,\,9\big)$	Integer
Ability estimate	$\hat{\theta} = (\mathrm{GII} - \alpha)/\beta$	Three decimals

Note. X *= number of the 157 items answered correctly;* $z = (\mathrm{GII} - 100)/15$*;* Φ *= standard normal distribution function;* α *and* β *are the equating coefficients of Table 5.1. The Scaled Score is clipped to the Wechsler subtest range, the GII is not, so the two ceilings differ (chapter 8). Rounding is to the nearest integer with halves rounded up. Omitted and not-reached items score 0 and are counted in* X*.*

One boundary in the table repays attention. The Scaled Score is clipped to the Wechsler range of 1 to 19 because that is the range the Wechsler scale defines; the GII is not clipped, because the deviation metric is unbounded. An examinee at the very top of the GIE therefore receives a Scaled Score of 19 and a GII near 150, and the Scaled Score, not the GII, is the score that has stopped discriminating. Where the two disagree, the GII is the one to read.

Alternative Conversions Examined

Four other conversions were computed on the linking sample and cross-validated against the observed criterion (Jouve, 2026b): a conditional-mean calibration; an item-level true-score equating routed through the WAIS-IN response functions; two distribution-matching equatings, an errors-in-variables linear equating (Deming, 1943) and a monotone-spline-smoothed equipercentile equating (Ramsay, 1988); and a correspondence-analysis person dimension. Through the range the WAIS-IN can measure, all of them coincide with the adopted equating, which is the reassurance a linking study is run to obtain (Kolen & Brennan, 2014).

They part company at the top, and the reason is instructive. The criterion subtest's hardest item sits at +1.24 logits, so the scaled scores of its ablest examinees censor near 16 to 17 whatever their true standing. A conversion fitted to those censored observations, or routed through the subtest's saturating response functions, inherits the censoring and caps a near-perfect GIE performance at the same place. The moment-anchored equating does not: it is anchored to the cohort's two moments and applied to the GIE's own item-level ability, so a score earned on items the subtest does not contain is credited rather than regressed toward the criterion's ceiling. The alternatives were set aside for that reason alone.

Scoring Under the Adaptive Form

The adaptive form under development changes item selection, not score meaning (van der Linden & Hambleton, 1997; Weiss, 2004). Because examinees no longer answer the same items, the raw score ceases to be interpretable and sufficiency no longer helps: ability must be estimated from the response pattern itself. The adaptive form therefore scores by the Expected A Posteriori (EAP) method under a standard normal prior, using the 2PL parameters of Table 7.3 (Bock & Mislevy, 1982; Thissen & Wainer, 2001). Omitted responses do not enter the likelihood, under a missing-at-random assumption:

$$p(\theta \mid \mathbf{x}) \;\propto\; \left[\prod_{i=1}^{I} P(X_i = 1 \mid \theta)^{x_i}\big(1 - P(X_i = 1 \mid \theta)\big)^{(1-x_i)}\right]\phi(\theta; 0, 1), \qquad \hat{\theta}_{\mathrm{EAP}} \;=\; \int \theta$$

The estimate is updated after each administered item and drives the selection of the next. At the end of the session the equating line $y(\hat{\theta}) = \alpha + \beta\hat{\theta}$ of Table 5.1 and the transformations of Table 5.2 produce the same report the fixed form produces: GII, WAIS-IN Scaled Score, T score, percentile, stanine, and $\hat{\theta}$. The two forms will therefore differ in how a score is obtained and not in what it means. The roadmap of chapter 3 commits to testing that claim with a linking study of its own before the two forms are treated as interchangeable.

CHAPTER 6

Reliability and Precision

This chapter reports the internal-consistency evidence for the 157-item fixed form, estimated on 821 examinees — the calibration records together with an additional set of operational administrations of the test, scored against the same key (Jouve, 2026c) — and translates that evidence into the score-level precision that individual interpretation requires. It reports no test–retest evidence, because none has been collected (chapter 9), and no conditional standard error, because the item parameters that would condition it are summarized in this manual rather than tabulated (Table 7.2).

Internal Consistency

The classical index KR-20 (Kuder & Richardson, 1937) equals .970 for the fixed form; on dichotomous items this coefficient coincides numerically with coefficient α (Cronbach, 1951) under the usual assumptions. Because any single division of a test into halves is arbitrary, reliability was additionally estimated with Jouve's Randomized Reliability Estimation (JRRE; Jouve, 2023, 2025), which

repeats the split-half computation over many random item partitions, applies the Spearman–Brown correction to each (Brown, 1910; Spearman, 1910), and summarizes the resulting distribution rather than a single arbitrary split.

The estimator ξ is the Monte-Carlo mean of the Spearman–Brown corrected split-half coefficients across the sampled partitions, and it is reported with two companions rather than alone (Jouve, 2025): the width of the 95% interval, which says how far the estimate depends on which items fall in which half, and the proportion of splits that meet the reliability standard the test's intended use requires. Across the random balanced partitions of the 821-examinee sample ξ = .970 (Jouve, 2026c). Table 6.1 reports it, and the same statistics for the criterion subtest against which the GIE was linked.

Table 6.1

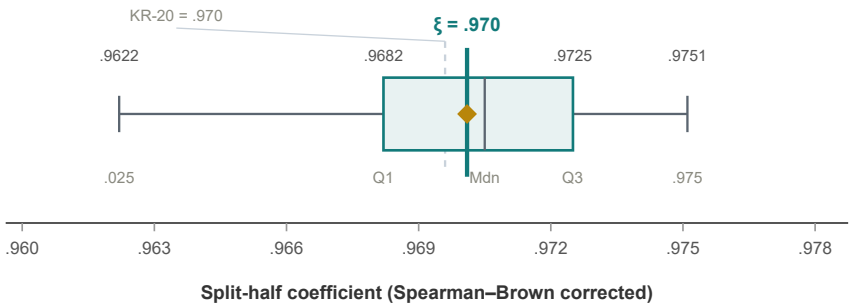
Internal-consistency reliability of the GIE and its criterion subtest

Instrument	Items	KR-20	ξ	95% interval	Width_{95}
GIE	157	.970	.970	[.962, .975]	.013
WAIS Information	25	.88	.88	—	—

Note. KR-20 = Kuder–Richardson formula 20 (Kuder & Richardson, 1937); ξ = the JRRE reliability estimate (Jouve, 2025), the Monte-Carlo mean of the Spearman–Brown corrected split-half coefficients across random balanced item partitions. For the GIE it was computed on 821 examinees — the calibration records and an additional set of operational administrations, scored against the same key (Jouve, 2026c). Each split correlated the two half-test totals (Pearson) and applied the correction $g(r) = \frac{2r}{1+r}$*; iteration stopped at* $\Delta\bar{\xi} < 10^{-9}$*.* Width_{95} *is the distance between the .975 and .025 quantiles of the split distribution, and the interval is those two quantiles; both describe dispersion across item partitions within this sample, not sampling error across samples. The linking study reports the criterion's ξ as a single value, without a distribution (Jouve, 2026b). That value, .88, is the criterion's reliability, not the disattenuated correlation of Table 7.1 (.90).*

Figure 6.1

Distribution of the randomized split-half coefficients



Note. The box spans the interquartile range, .9682 to .9725; the whiskers mark the .025 and .975 quantiles, the 95% interval of Table 6.1. The bold line marks ξ = .970, the mean of the split distribution; the thin interior line marks the median (.9705); the dashed line marks KR-20 = .970 (Kuder & Richardson, 1937). Skewness = −0.773.

Figure 6.1 summarizes the split distribution. The negative skew, a longer tail of lower coefficients, is the expected shape for a statistic bounded above by 1.00, and the entire spread across the sampled partitions occupies about one hundredth of the reliability scale.

Read against the thresholds the method paper sets, the evidence is unambiguous (Jouve, 2025). The coefficient of .970 clears .90, the standard reserved for high-stakes individual decisions, with a great deal to spare. The 95% interval is .013 wide, where any width at or below .05 counts as very stable across splits. And its lower bound, .962, lies so far above the standard that every sampled partition meets it $\left(\pi_{.90} = 1.000\right)$. Whether the GIE is reliable enough for individual interpretation therefore does not depend on which items happen to fall in which half, which is the question a single split-half coefficient cannot answer and the reason for computing many of them.

That the classical coefficient and the randomized one agree is expected rather than fortunate. Classical indices are accurate when items are well behaved, when the data are complete, and when the form is long, which the method paper puts at $k \geq 40$ items for heterogeneous content. The GIE has 157 items across five content domains, a mean discrimination of 1.306 (Table 7.3), and no missing-data mechanism, since an unanswered item is scored as failed rather than dropped. Nor does any of the five conditions that call for a caution note apply: the form is not short, the interval is not wide, the coefficient is nowhere near

a decision threshold, no item discriminates below .5 (the least discriminating sits at 0.686), and the responses are dichotomous rather than nominal. What the randomized estimate contributes here is therefore not a better point estimate, which KR-20 already supplies, but the statement of uncertainty that a coefficient computed once cannot.

Length is what separates the GIE from its criterion. The Spearman–Brown relation (Brown, 1910; Spearman, 1910) gives the reliability a form of a different length would show, provided the items added or removed are essentially τ-equivalent to those retained. For a length factor k relative to the 157-item form,

$$r'(k) \;=\; \frac{k\,r}{1+(k-1)r}, \qquad r = 0.970, \qquad r'\!\left(\tfrac{25}{157}\right) \approx 0.84.$$

A randomly chosen 25 of the GIE's items, about a sixth of the form, would therefore be expected to reach ξ ≈ .84. The 25-item WAIS Information subtest reaches .88 (Table 6.1). The criterion is thus a well-built short test rather than a weak one: its items were selected to discriminate, and it carries a discontinue rule the GIE does not use. What separates the two instruments is length, and the subtest makes better use of the length it has than a random slice of the GIE would.

The same assumption is why the relation cannot be turned around to forecast the adaptive form. Spearman–Brown describes a form shortened at random; a computerized adaptive test selects, for each examinee, the items that carry the most information about that examinee's ability. Its precision is therefore governed by the test information function and not by length alone, and a projection computed from the fixed form's coefficient neither bounds it from above nor from below. Parity between the two forms is a claim to be established by the adaptive form's own linking study (chapter 3), not projected here.

Score-Level Precision

On the GII metric, this reliability translates into measurement error as follows:

$$\text{SEM} = \sigma_X\sqrt{1-\xi} \quad \Rightarrow \quad \text{SEM} = 15 \times \sqrt{1-0.970} \approx \mathbf{2.60}\ \text{GII}, \qquad 95\%\ \text{score band}$$

A reliability coefficient is a property of a sample of scores; a standard error of measurement is what an individual report has to carry. With a GII-scaled *SEM* of about 2.6 points, a reported GII is accurate to within about five points at the conventional level of confidence, or one Scaled Score point on the criterion metric. That precision supports fine-grained individual interpretation, provided scores are reported with their confidence bands and read against the validity evidence of chapter 7. The bands themselves, and the differences between scores that they do and do not permit, are the subject of chapter 8.

CHAPTER 7

Validity Evidence

The GIE's validity argument rests on three lines of evidence, corresponding to the *Standards*' categories of evidence based on relations to other variables and on internal structure (AERA, APA, & NCME, 2014): agreement with the WAIS Information subtest as an external criterion, the fit of the calibration that carries GIE scores onto that criterion's metric, and structural evidence, within the bank and across both instruments, that a single dimension is being measured.

Criterion Agreement With WAIS Information

Table 7.1 collects the three coefficients the criterion study reports, each a different answer to the question of how closely the two instruments agree.

Table 7.1

Agreement between the GIE and the WAIS Information subtest

Comparison	*N*	*r*	95% CI	RMSE	MAE
Observed scores, deviation metric	482	.816	[.79, .84]	8.99	6.94
Rasch ability estimates	483	.82	—	—	—
Rasch ability estimates, disattenuated	483	.90	—	—	—

Note. Row 1 is computed on the deviation-IQ metric obtained from WAIS-IN Scaled Scores by $\mathrm{IQ} = 100 + 15(\mathrm{SS} - 10)/3$*, a linear transform that leaves the correlation unchanged; its interval follows the Fisher r-to-z transformation given in the text. RMSE = root-mean-square error and MAE = mean absolute error of the fitted conversion, in IQ points. Rows 2 and 3 are from the concurrent Rasch calibration of the linking study, which analyzed 483 adults; the disattenuated value corrects the observed agreement for the two ability estimates' empirical reliabilities (Jouve, 2026b; Spearman, 1910). Dashes mark statistics the study does not report for those rows.*

The three rows answer three different questions, and only the first is about the instruments as they are used. Observed scores agree at r = .816, and the fitted conversion misses the criterion by about 9 IQ points in root-mean-square and about 7 in mean absolute error, which is between one and two Scaled Score points, and a useful figure to hold in mind when a GIE result is set beside a WAIS Information result for the same person. The second row removes the conversion from the comparison and correlates the two ability estimates directly. The third removes measurement error from both instruments and asks what the agreement would be between two perfectly reliable measures of whatever each is measuring. It is the third that bears on the construct question, and it is the third that chapter 5 relies on. The confidence interval for the first derives from the Fisher transformation:

$$z = \tfrac{1}{2}\ln\frac{1+r}{1-r}, \qquad \mathrm{SE}_z = \frac{1}{\sqrt{N-3}}, \qquad z_{\pm} = z \pm 1.96\,\mathrm{SE}_z, \qquad r_{\pm} = \frac{e^{2z_{\pm}}-1}{e^{2z_{\pm}}+1}.$$

With $r = .816$ and $N = 482$, the interval is approximately $[.79,\ .84]$. The observed coefficient understates the latent agreement between the instruments. In the concurrent Rasch calibration of the linking study

(Jouve, 2026b; Rasch, 1960), ability estimates from the two item sets correlated .82, and correcting for the two ability estimates' empirical reliabilities raised the disattenuated correlation to approximately .90, near the value expected when two tests estimate one latent ability, with the residual gap consistent with the subtest's 25-item length and discontinue scoring rather than with construct differences. Because the linked sample was self-selected and over-dispersed relative to the WAIS normative population, these coefficients are best read as upper bounds (Jouve, 2026b; Sackett & Yang, 2000).

Fit of the Conversion

Against the observed criterion, the equating function of chapter 5 accounts for $R^2 = .631$ of the variance in WAIS-IN performance, with residual error on the IQ metric of RMSE ≈ 9 points and MAE ≈ 7 points, and it required no extrapolation anywhere in the observed ability range (θ coverage −3.0649 to 3.3534). Being linear with a positive slope, it is strictly increasing, so score order is preserved along the entire continuum.

That R^2 is below the square of the correlation in Table 7.1, and the gap is a property of the conversion rather than a shortcoming of it. A least-squares line through the same points would account for more variance, because that is what least squares minimizes; it would do so by shrinking the predicted scores toward the criterion mean, and a conversion that shrinks the spread of the scale it produces is not a conversion between score scales. The moment-anchored line preserves the criterion's standard deviation and pays for it in residual variance. The residual is what Table 7.1 reports as RMSE, and it is the quantity to cite when a GIE result and a WAIS Information result for one examinee do not agree.

One Dimension Across Both Instruments

A criterion correlation alone cannot show that two tests measure the *same* dimension rather than correlated neighbors. The linking study (Jouve, 2026b) addressed dimensionality directly on 483 adults who completed both instruments, analyzing a combined pool of 178 items

(154 GIE and 24 WAIS-IN items retained after setting aside items without usable variance). A NOHARM normal-ogive item factor analysis (Fraser & McDonald, 1988) fitted a single latent dimension almost exactly (residual RMS = .011, Tanaka goodness-of-fit index = .999), and a second dimension improved the fit only trivially (RMS = .009); items from the two sources loaded comparably on the shared dimension (mean loadings .58 for the GIE, .61 for the WAIS-IN). Multidimensional scaling of the inter-item proximities displays this structure geometrically; Table 7.2 and Figure 7.1 report it, and the limits on what such a display can establish are stated below (Jouve, 2026a).

The scaling was performed by alternating least squares (Takane, Young, & de Leeuw, 1977) on the matrix of inter-item simple-matching proximities, under the interval model. Table 7.2 gives the adequacy of the resulting configuration at one through six dimensions. Beyond two dimensions the improvement in the representation of the proximities is small: the reduction in Stress halves at each further dimension, and the additional variance accounted for is bought at the cost of an axis that admits no interpretation.

Table 7.2

Adequacy of the item configuration by dimensionality

Solution	Stress	R.S.Q.
1 dimension	0.364	0.833
2 dimensions	0.236	0.847
3 dimensions	0.173	0.886
4 dimensions	0.138	0.908
5 dimensions	0.116	0.925
6 dimensions	0.101	0.936

Note. ALSCAL of the inter-item simple-matching proximities, interval model, on the combined pool of 178 items answered by 483 adults (Jouve, 2026b). Stress is Kruskal's formula 1, a badness-of-fit index: it falls as dimensions are added, and reaches 0 for a configuration that reproduces the proximities exactly. R.S.Q. is the

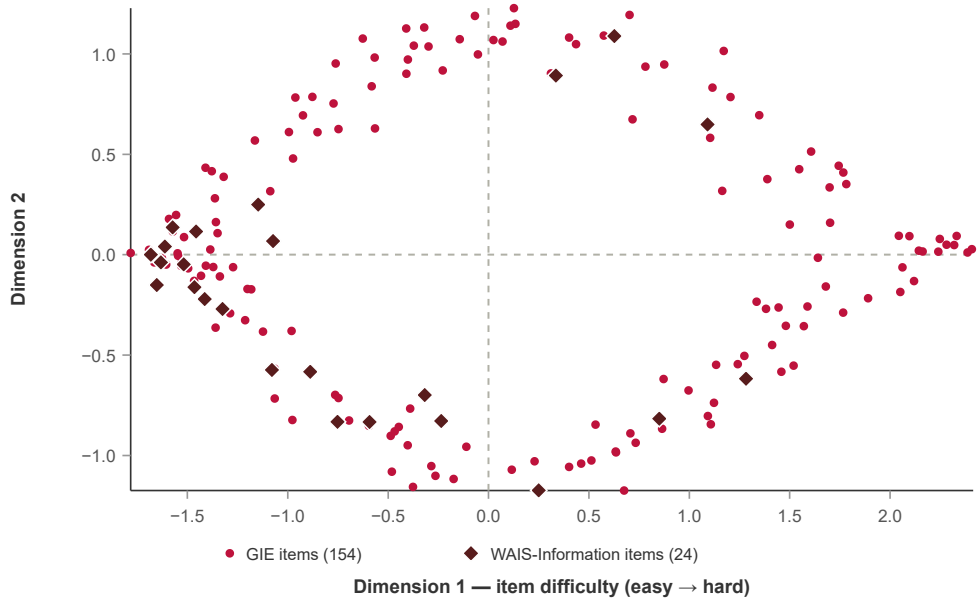
proportion of variance in the scaled proximities accounted for by their distances in the configuration.

Figure 7.1 shows the two-dimensional solution. Dimension 1 is difficulty: it correlates 1.00 with the proportion of examinees answering an item correctly, so the items run from easy at the left of the figure to hard at the right. Dimension 2 is not a second ability, and it is not the horseshoe curvature that item maps are known for: that fold is the signature of a proximity that stops discriminating beyond a local horizon — the mechanism of Diaconis, Goel, and Holmes (2008) — and simple matching, which remains informative across the whole difficulty span, produces none of it on this pool (Jouve, 2026a). What the second dimension carries is the one structure the item associations hold beyond difficulty: a science–humanities facet internal to crystallized knowledge that cuts across both instruments alike. That facet is why the items describe a ring rather than a single band, the sciences arching above and the humanities below.

What the configuration can establish deserves care, because a companion study fixes its limits: agreement-based proximities order items by their difficulties even when no shared dimension exists, so the interleaving of the two instruments along dimension 1 would appear whether or not they measured the same thing (Jouve, 2026a). The dimensionality conclusion of this section therefore rests on the factor analysis and the concurrent calibration above; the configuration illustrates it rather than adding to it. Where the figure does carry independent information is the second axis: the only association-borne structure beyond difficulty is the content facet, and it divides the sciences from the humanities — on both instruments alike — rather than the GIE from the criterion, with item source accounting for 1% of the variance on that axis. One absence is also informative: the diamonds stop before the right-hand end, where the hardest items are the GIE's alone; that is the subject of the next paragraph.

Figure 7.1

Configuration of the 178 items in a two-dimensional space

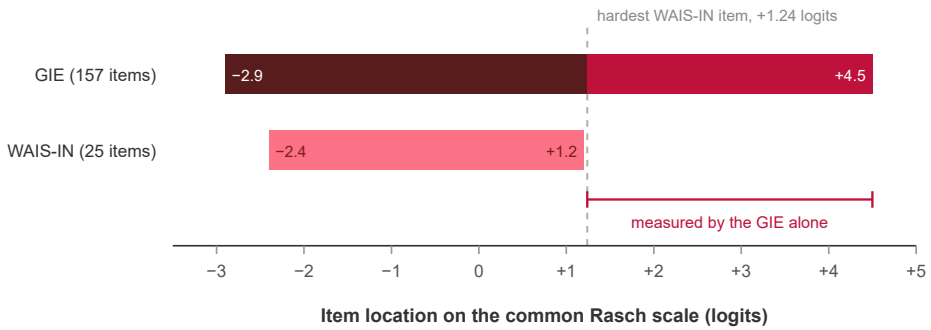


Note. Item configuration from the two-dimensional ALSCAL scaling of the inter-item simple-matching proximities, interval model, on the 483 adults who took both instruments (Jouve, 2026b; Takane, Young, & de Leeuw, 1977). Table 7.2 gives the fit. Dimension 1 is item difficulty: it correlates 1.00 with the proportion answering correctly, so the items are ordered from easy at the left to hard at the right. Because simple matching remains informative across the whole difficulty range, the configuration shows none of the horseshoe fold that locally informative proximities produce (Jouve, 2026a; Diaconis, Goel, & Holmes, 2008); the ring's two arcs are instead the science–humanities content facet, which cuts across both instruments alike and carries the only association-borne structure beyond difficulty. Item source explains 1% of the variance on that axis. The two instruments interleave around the whole figure, and the diamonds stop before the right-hand end: the hardest items are the GIE's alone.

The same study placed both item sets on a common Rasch scale (Rasch, 1960). Figure 7.2 puts the two item sets on it side by side. GIE item locations spanned −2.9 to +4.5 logits, whereas the 25 WAIS-IN items occupied a band from −2.4 to +1.2 logits that falls entirely within the GIE range, the subtest's hardest item sitting at +1.24 logits. The reference subtest is thus, in effect, a short and easier slice of the continuum the GIE measures end to end, and the GIE resolves abilities above the subtest's ceiling. The only structure recovered beyond difficulty was a minor science–humanities content tilt on the second scaling axis (two-way $\eta^2 = .50$); it was orthogonal to difficulty and cut across both instruments alike: an internal facet of crystallized knowledge, not a distinction between the tests (Jouve, 2026b).

Figure 7.2

Item difficulty coverage of the two instruments on a common Rasch scale



Note. Item locations from the concurrent Rasch calibration of the linking study (Jouve, 2026b; N = 483). The 25 WAIS-IN items occupy a band that falls entirely inside the range the 157 GIE items cover. The subtest's hardest item sits at +1.24 logits (dashed line); the shaded remainder of the GIE bar is the part of the continuum on which the criterion has no items and cannot separate examinees. Bar extent is the range of item locations, not a confidence interval.

Item-Parameter Evidence

Within the GIE bank itself, the 2PL calibration indicates well-behaved items with adequate discrimination and broad difficulty coverage. Table 7.3 summarizes the item parameters.

Table 7.3

Two-parameter logistic item parameters of the operational bank

Parameter	*M*	*SD*	Minimum	Maximum
Discrimination, a	1.306	0.262	0.686	2.130
Difficulty, b	0.002	1.259	−2.093	4.000

Note. Two-parameter logistic calibration of all 157 operational items, on the standard normal θ metric. Discrimination governs how sharply an item separates neighbouring abilities; difficulty locates the ability at which an item is answered correctly half the time.

Item difficulties center almost exactly on the ability mean (M = 0.002) and disperse widely around it (SD = 1.259), so measurement information is available from low to high general-knowledge proficiency rather than concentrated where examinees are most numerous. Discriminations average 1.306, and the least discriminating item in the bank sits at 0.686 (Table 7.3), low but not so low that the item contributes length without contributing precision. The difficulty range extends to 4.000, four standard deviations above the mean of the ability distribution, and it is those items that give the form the reach Figure 7.2 displays.

Two limits on this chapter should be read with it. All criterion evidence is sample-based and specific to one criterion, the WAIS Information subtest; agreement with a single instrument, however close, establishes convergent validity within a narrow neighbourhood of the construct and says nothing about discriminant validity against the abilities the GII is not meant to measure. And the linked sample was self-selected. Both bear on how far the coefficients above should be generalized, and chapter 9 states the consequences.

CHAPTER 8

Score Interpretation

The GII is interpreted against the population: on the deviation metric each 15-point step marks one standard deviation from the reference mean, and T-scores, percentiles, stanines, and WAIS-IN Scaled Scores express the same standing in alternative units. Because the GIE is equated to the WAIS Information reference, one Scaled Score point equals 5 GII points (a Scaled Score of 12 corresponds to GII ≈ 110). Table 8.1 gives the population-referenced classification bands.

Table 8.1

Classification ranges (population-referenced)

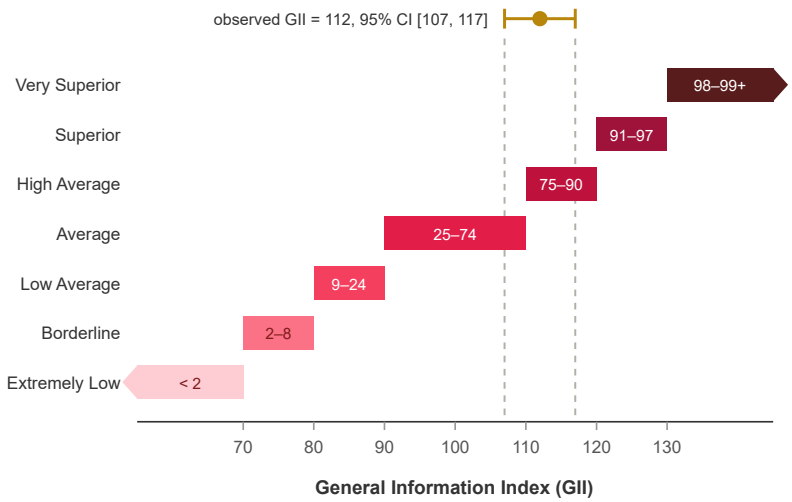
GII Range	Classification	Approx. Percentile Range	Typical Description
130+	Very Superior	98–99+	Exceptionally broad general knowledge
120–129	Superior	91–97	Well above typical knowledge breadth
110–119	High Average	75–90	Above-typical factual knowledge
90–109	Average	25–74	Typical knowledge for the reference population
80–89	Low Average	9–24	Somewhat below typical breadth of knowledge
70–79	Borderline	2–8	Limited knowledge across several domains
< 70	Extremely Low	< 2	Marked limitations in factual knowledge

Note. GII = General Information Index (M = 100, SD = 15). Percentile ranges follow the normal model applied to the GII metric. Classifications describe score ranges, not persons; a score near a band boundary should be classified with its confidence band (see below).

The bands are conventions, not discoveries. Nothing in the data distinguishes a GII of 109 from a GII of 110 beyond the one point that separates them, yet the first is labelled Average and the second High Average. The labels are useful for communicating a standing quickly and dangerous for anything else, because a boundary invites a reader to treat a difference of one point as a difference of kind. Figure 8.1 is drawn to make that danger visible.

Figure 8.1

Classification bands with the confidence band of an observed score



Note. Bands and percentile ranges are those of Table 8.1 on the GII metric (M = 100, SD = 15); the label inside each bar is its approximate percentile range. The gold interval is the 95% confidence band of a single observed GII of 112 (±1.96 SEM, SEM = 2.6), and the dashed verticals carry its limits down through the bands: the score is classified High Average, but its interval reaches into Average. A score this close to a boundary is classified by its band and its collateral evidence, not by its point estimate. The lowest and highest bands are open-ended; bars are truncated at GII 55 and 145 for display.

Measurement Precision and Confidence Bands

Every reported score is an estimate, and interpretation should carry the corresponding band. Applying the JRRE reliability on the GII metric gives the standard error of measurement:

$$\mathrm{SEM}_{\mathrm{GII}} \;=\; SD \times \sqrt{1-\xi} \;=\; 15 \times \sqrt{1-0.970} \;\approx\; 2.60, \qquad 95\%\,\mathrm{CI}: \quad \mathrm{GII} \pm 1.96 \times$$

On the WAIS-IN Scaled Score metric the *SEM* is one-fifth as large (1 Scaled Score point = 5 GII points): $\mathrm{SEM}_{\mathrm{SS}} \approx 0.50$, giving a 95% interval of roughly ±1 Scaled Score point. In practice, an observed GII of 112 carries a 95% interval of about 107–117 (Scaled Score 12, interval 11–13), and a GII of 88 about 83–93 (Scaled Score 8, interval 7–9). These bands describe the expected imprecision of a single administration and belong in any individual decision.

Figure 8.1 lays the first of those intervals across the classification bands. The point estimate of 112 falls in High Average, two points above the boundary; the interval reaches from 107, which is Average, to 117, which is not. The examinee's standing is not in serious doubt (it lies somewhere between the 68th and 87th percentile), but the label is, and a report that gives the label without the interval has converted an uncertainty about words into an apparent certainty about ability. Where a score sits near a boundary, the band and the collateral evidence carry the classification, not the point estimate.

Reading GIE Results on the WAIS-IN Metric

Table 8.2 anchors the correspondence among Scaled Scores, GII values, and percentiles. The relation between the two score metrics is the linear identity of Table 5.2, so the table is a lookup convenience rather than a finding; its upper rows are the exception. There the GIE reports a standing the criterion subtest cannot: the WAIS Information subtest has no items above +1.24 logits and cannot separate examinees who pass everything it contains, whereas the GIE's items extend to +4.5 (Figure 7.2), and the moment-anchored conversion of chapter 5 credits that ability rather than regressing it toward the criterion's censored ceiling (Jouve, 2026b).

Table 8.2

WAIS-IN Scaled Score to GII correspondence

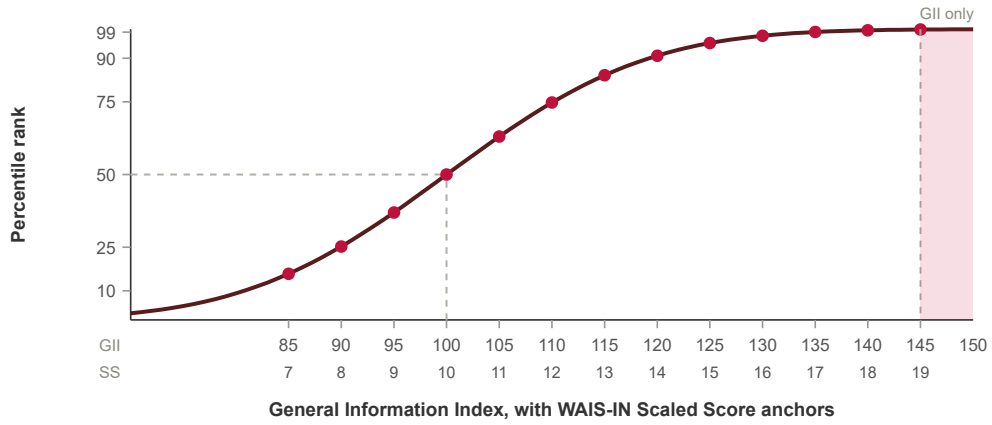
WAIS-IN SS	Approx. GII	Percentile (approx.)
7	85	~16th
8	90	~25th
9	95	~37th
10	100	50th
11	105	~63rd
12	110	~75th
13	115	~84th

WAIS-IN SS	Approx. GII	Percentile (approx.)
14	120	~91st
15	125	~95th
16	130	~98th
17	135	~99th
18	140	> 99th
19	145	> 99.9th

Note. Correspondence follows $\mathrm{GII} = 100 + 5(\mathrm{SS} - 10)$*; percentiles follow the normal model and are approximate. SS = WAIS-IN Scaled Score (M = 10, SD = 3). The uppermost rows lie beyond the criterion subtest's effective ceiling and rest on the GIE's own item statistics (Jouve, 2026b).*

Figure 8.2

Percentile rank as a function of the GII



Note. The curve is the normal model, percentile = 100 × Φ[(GII − 100)/15]. Points mark the thirteen anchor conversions of Table 8.2; the two rows beneath the axis give each anchor's GII and its WAIS-IN Scaled Score. The mapping between the two score metrics is linear by construction, so the curvature here is entirely the percentile's. Equal steps in GII buy large percentile gains near the mean and small ones at either tail, which is why a percentile difference is a poor measure of an ability difference. The shaded strip beyond GII 145 is the range the GIE reports and the Scaled Score, capped at 19, cannot express (chapter 5).

Figure 8.2 plots the third column of that table against the first, and the curve is the reason the third column should be quoted with care.

Between GII 95 and GII 105, a distance of two Scaled Score points, the percentile rank moves by about 26 points. Between GII 130 and GII 140, the same two Scaled Score points move it by about two. Percentile rank is a statement about how many people a score exceeds, not about how much ability separates them, and near the tails the two diverge sharply. Score differences are compared on the GII, where the unit is constant; percentiles are for explaining a standing to a reader who is not going to compare anything.

The shaded strip at the right of the figure is the range in which the two metrics part company. A Scaled Score cannot exceed 19, because the Wechsler subtest scale stops there; the GII can, and at the maximum raw score it reaches about 150. An examinee in that strip has answered items the criterion subtest does not contain, and the GII is the only one of the two scores still carrying information about how far above the ceiling the performance lies.

Meaningful Differences Between Scores

When two administrations of the same individual are compared, for progress monitoring or for verification, the relevant error is not the standard error of either score but the standard error of their difference, $\mathrm{SE}_{\Delta} \approx \sqrt{2} \times 2.60 \approx 3.68$ on the GII scale. A change must therefore exceed about $1.96 \times 3.68 \approx 7$ GII points, or roughly 2 Scaled Score points, before it can be attributed to anything other than measurement error. A gain of 5 points between two administrations is, on this evidence, no gain at all.

The figure is a lower bound on what a real change must clear, because it counts only measurement error. Practice effects, the interval between administrations, and regression toward the mean for scores selected because they were extreme all move a second score independently of ability, and none of them is captured by SE_{Δ}. A formal reliable-change index requires test–retest data, which the GIE does not yet have; until it does, the interval above should be treated as the smallest difference worth discussing rather than as the threshold at which a difference has been established.

Interpretive Considerations

Three disciplines keep interpretation sound. First, the point estimate and its 95% interval travel together; a GII reported without its band overstates what one administration can establish. Second, the GIE measures learned knowledge, so educational history, language exposure, and access to information belong in every reading of a score, particularly when groups or individuals with different opportunities are compared. Third, the GII summarizes crystallized knowledge and nothing else; when the referral question extends to fluid reasoning, processing speed, or memory, the GIE should be paired with instruments that measure those domains (Schneider & McGrew, 2018).

CHAPTER 9

Appropriate Use and Limitations

The GIE is a validated fixed-form measure of crystallized general knowledge, suitable for educational and occupational contexts in which breadth of acquired information is the question, and for research on the Gc domain. Its evidence base is strong but bounded, and responsible use requires stating where the boundaries lie.

Two boundaries follow from what has not yet been measured. No formal test–retest study has been conducted, so the manual can say what a single administration establishes but not what two establish; until such a study exists, a small change over a short interval should be judged against the confidence band of chapter 8 rather than at face value, and the reliable-change reasoning offered there treated as a lower bound. And scores are trustworthy over the range of performance the calibration actually observed. That range is wide, spanning very nearly the whole of the fixed form (chapter 3), but a result at either extremity rests on fewer observations than one near the mean, and should be reported as beyond the range the calibration exercised.

The sample on which the linking rests was self-selected and, on the criterion, unusually able: its WAIS-IN Scaled Scores averaged 12.7 (*SD* =

3.0) where the normative mean is 10. Two consequences follow. Absolute associations computed in a sample more dispersed than the reference population are inflated, so the validity coefficients of chapter 7 are best read as upper bounds (Jouve, 2026b; Sackett & Yang, 2000). And the conversion of chapter 5 is anchored to that sample's criterion distribution rather than to fresh census-referenced norms, which is why its intercept sits above 100. A GII is therefore a standing relative to the linking cohort as it was constituted, expressed on a metric the reader will recognize, and not a standing relative to a census-representative population.

The GIE measures learned knowledge, and education, language exposure, and access to information legitimately influence what a person has had the opportunity to learn. This is a property of the construct, not a defect of the instrument, and it is why a GII is evidence about crystallized knowledge rather than about the capacity that produced it. Group fairness analyses, including differential item functioning review, are underway and are not reported in this edition; until they are, comparisons across groups that differ in educational opportunity should be interpreted in the light of that difference and not in isolation from it.

Administration mode separates the GIE from the instrument it was linked to. The GIE is taken online and unsupervised; the criterion data were collected under proctored, standardized administration. Mode differences contribute construct-irrelevant variance, and for general-knowledge content delivered online they include the possibility of unsupervised answer look-up, against which no scoring procedure protects. The equating accordingly links an online score to a metric defined under standardized conditions (Jouve, 2026b). For consequential decisions a quiet, supervised setting is preferred and irregularities should be recorded. A related caution attaches to the adaptive form: until its linking study is published, fixed-form and adaptive scores should not be combined in one decision without noting which form produced which, and follow-up testing should match the conditions of the original administration as closely as it can.

Finally, the GII summarizes crystallized knowledge and nothing else. It does not measure fluid reasoning, processing speed, working memory,

or clinical status, and it is not suitable as a sole measure for diagnosis (AERA, APA, & NCME, 2014); when the referral question extends beyond acquired knowledge, other instruments must carry it. Precision also tapers toward the tails, and the uppermost conversions rest on the GIE's own item statistics beyond the criterion subtest's range rather than on criterion observations (Jouve, 2026b). That is a strength of the instrument, since the criterion cannot resolve those abilities at all, and simultaneously the point at which its scores are least anchored to anything external. Extreme scores warrant confidence-band reasoning and a narrative explanation, never a single number.

The development roadmap addresses these boundaries in the order in which they bind. A test–retest study with reliable-change guidance would close the first. Validation against external criteria beyond the WAIS Information subtest, and against criteria drawn from populations less selected than the linking cohort, would address the second and third together. The fairness work already underway, differential item functioning review and content refresh, addresses the fourth. And the adaptive form will be released only with documented linking to the metric this manual describes. Each of these is a commitment of the development programme, and none of them is evidence; where this manual reports no evidence, none should be assumed.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Addison-Wesley.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6*(4), 431–444. https://doi.org/10.1177/014662168200600405

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3*(3), 296–322. https://doi.org/10.1111/j.2044-8295.1910.tb00207.x

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge University Press.

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology, 54*(1), 1–22. https://doi.org/10.1037/h0046743

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334. https://doi.org/10.1007/BF02310555

Deming, W. E. (1943). *Statistical adjustment of data*. Wiley.

Diaconis, P., Goel, S., & Holmes, S. (2008). Horseshoes in multidimensional scaling and local kernel methods. *The Annals of Applied Statistics, 2*(3), 777–807. https://doi.org/10.1214/08-AOAS165

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates.

Fraser, C., & McDonald, R. P. (1988). NOHARM: Least squares item factor analysis. *Multivariate Behavioral Research, 23*(2), 267–269. https://doi.org/10.1207/s15327906mbr2302_9

Jouve, X. (2023). *Jouve's randomized reliability estimation (JRRE) calculator*. Cogn-IQ Statistical Tools. https://pubscience.org/ps-1mSLX-obf95e-ik5S

Jouve, X. (2024). *General Information Evaluation (GIE)*. Cogn-IQ Cognitive Assessments. https://pubscience.org/ps-1mVDO-38a233-oHvW

Jouve, X. (2025). When alpha fails: Jouve's randomized reliability estimation (ξ) versus classical reliability coefficients in Rasch, 2PL, 3PL, 4PL, GRM, GPCM, and NRM models. *Cogn-IQ Research Papers*. https://pubscience.org/ps-1mYdi-014f7f-ormI

Jouve, X. (2026a). The difficulty horseshoe in item-level multidimensional scaling as a signature of local information. *Cogn-IQ Research Papers*. https://pubscience.org/ps-1n0JE-2ebae9-MqDI

Jouve, X. (2026b). Linking the General Information Exam to WAIS Information: One crystallized-knowledge dimension and a raw-to-scaled conversion. *Cogn-IQ Research Papers*. https://pubscience.org/ps-1n0JH-be9993-znpb

Jouve, X. (2026c). Split-invariant reliability of the General Information Evaluation: A randomized reliability estimate for the General Information Index. *Cogn-IQ Research Papers*. https://pubscience.org/ps-1n0JV-4f433d-yJ25

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). Springer.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*(3), 151–160. https://doi.org/10.1007/BF02288391

McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence, 37*(1), 1–10. https://doi.org/10.1016/j.intell.2008.08.004

Ramsay, J. O. (1988). Monotone regression splines in action. *Statistical Science, 3*(4), 425–441. https://doi.org/10.1214/ss/1177012761

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Danmarks Pædagogiske Institut.

Sackett, P. R., & Yang, H. (2000). Correction for range restriction: An expanded typology. *Journal of Applied Psychology, 85*(1), 112–118. https://doi.org/10.1037/0021-9010.85.1.112

Schneider, W. J., & McGrew, K. S. (2018). The Cattell-Horn-Carroll theory of cognitive abilities. In D. P. Flanagan & E. M. McDonough (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (4th ed., pp. 73–163). Guilford Press.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3*(3), 271–295. https://doi.org/10.1111/j.2044-8295.1910.tb00206.x

Takane, Y., Young, F. W., & de Leeuw, J. (1977). Nonmetric individual differences multidimensional scaling: An alternating least squares method with optimal scaling features. *Psychometrika, 42*(1), 7–67. https://doi.org/10.1007/BF02293745

Thissen, D., & Wainer, H. (Eds.). (2001). *Test scoring*. Lawrence Erlbaum Associates.

van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. Springer.

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale–Fourth Edition (WAIS-IV)*. Pearson.

Weiss, D. J. (2004). Computerized adaptive testing for effective and efficient measurement in counseling and education. *Measurement and Evaluation in Counseling and Development, 37*(2), 70–84. https://doi.org/10.1080/07481756.2004.11909751